ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20171202222013
Original scientific paper

Development of a Data Mining System for Subscriber Classification (Case Study: Electricity Distribution Company)

Elham PEYK, Asadollah SHAHBAHRAMI

Abstract: Currently, organizations and companies tend to provide customers with good and suitable services in accordance with their interests and behaviors. Thus, the better the customers are classified, the better the services provided will be. Data mining is an efficient process for helping companies discover patterns in the database and it is important to identify target customers in this process. In fact, customers are selected to provide new products and services. Customer classification is based on data mining techniques for customer identification. This study tends to classify customers using data mining algorithms such as decision tree CART, neural network and regression. The case study is customers of Electricity Distribution Company. Simulation results based on Clementine software show that population had the highest effect on the amount of power consumed in each of the six household, public, industrial, agricultural, road and commercial classes. This is consistent with the opinion of experts in the electric power industry, because higher number of subscribers of each class surely increases the amount of electricity consumed (not steadily). The second effective feature of power consumption in six classes is humidity, which in many classes has a relatively equivalent effect with the effect of temperature on power consumption.

Keywords: classification; customers; data mining; power consumption

1 INTRODUCTION

The source of current profit and growth of an organization is customer. However, it is difficult to recognize, attract and retain a good customer who is profitable for the organization due to the increased awareness of customers and, consequently, the change in their level of expectation and the existence of close competition. This is partially possible with proper utilization of information technology, customer identification and management. Throughout the world, organizations do not treat their customers equally; by rating customers, they offer people who are rated better some special benefits. Some organizations use other methods to classify customers. Data mining is a new and powerful technique to help companies discover patterns in customer database. Direct marketing process is related to attracting customers and it is very important to identify target customers in this process. In fact, target customers are selected to provide new products and services. Therefore, it is vital to identify target customers in a competitive environment. Customer classification is based on data mining techniques to identify target customers and when the number of records is high. However, it does not yield a good result when the number of records is low [1].

Classification of subscribers is based on their type of activity. For example, the electricity industry classifies subscribers to determine electricity tariffs. Effective factors on structure of electricity tariffs can be divided into two categories of internal and external factors. Internal factors refer to how electricity is supplied and external factors refer to subscribers. Meanwhile, the role of external factors is particularly important. Different models and different methods are used for classifying subscribers. For analysis of domestic power consumption, for example, domestic power subscribers are classified into low-consumption and high-consumption subscribers. In another analysis, all subscribers are classified according to their needs and type of use in different groups. Statistics of the electricity industry shows the amount of consumption for each subscriber individually; however, its analysis requires different methods and, given the high volume of data used, data mining tools can be useful. Previous works related to consumption were related to annual consumption and consumption forecast; however, fewer works have been done for analysis and classification of subscriber consumption per month. Taking into account monthly subscriber consumption, this study tends to classify them in different groups [2, 3].

Mirdehghan and Saadatjoo [4] classified subscribers to optimize their consumption. This study tended to analyze and predict power consumption using data mining techniques and obtain the effect of different factors on consumption. They used the prediction to manage and optimize the power consumption. This study used five data mining algorithms including MLP3, SVR, random forest, classification and stepwise regression. The performed studies show that SVR algorithm has better results than other algorithms.

Long-term prediction of electricity consumption in Greece was made using neural networks. MLP method was used for prediction. The network input included data collected from 1992 to 2008. The goal was to predict electricity consumption between 2009 and 2015. Relative error of the estimated values with actual values was 2% [5, 6].

Through a study on prediction of power consumption using linear regression, neural networks and decision tree in Hong Kong, it was shown that addition of factors such as air temperature and wind velocity improved the suggested model. By comparing output of the models, it was concluded that decision tree and neural networks, respectively, had better predictions [7].

Clustering technique was used for domestic electrical data and electricity charge profile; it was found that the technique used in Portugal (a two-step process involving self-organizing maps and K-means) could not be used for British information [8]. Research has shown that more than 9 clusters of domestic consumers could be distinguished which differed in profiles. This shows that the pattern of power consumption could be clustered to more than two groups. The data was based on findings from the Milton Keynes area in 1990. Further work is done on making significant and accurate clusters of similar electricity

consumers in order to better manage sample groups and similar subscribers [9].

In Turkey, optimal power consumption was modelled based on population, gross domestic product, imports and exports. The prediction done in this study was related to 2007-2026. According to predictions, electricity consumption will increase to 284 terawatt/hour in 2026 [10]. Short-term prediction of power consumption was done by gray method for the member countries of the Asia-Pacific Economic Cooperation Organization. This theory applies when information is inadequate and unknown. When the dataset is low, the gray model can be useful. The results show that the mean absolute error of the prediction model is less than 3.1% [11]. Kaur and Kaur [12] used Artificial Neural Networks to predict electricity consumption. Effective factors were temperature, humidity and public holiday. Abolfazli et al. [13] predicted rail transport petroleum consumption using an autocorrelation functions artificial neural network. Dalfard [14] used a data mining techniques based adaptive network-based FIS estimate and forecast both long-term electricity and natural gas (NG) consumptions. Kuo et al. [15] applied big data mining, machine learning techniques to predict convenience stores energy consumption performance. Key factors were lighting and refrigeration, and relative humidity.

This study tends to compile a dataset including the amount of electricity consumed in recent years in one of the provinces of Iran and independent statistics (factors involved in the amount of electricity consumed) such as population, humidity, air temperature and power consumption. This data is collected from the Statistical Center, Tavanir, and Meteorological Center monthly and annually from 1994 to 2014. Using data mining techniques, an analysis is performed on this dataset, and the power consumption model is extracted from existing data [16].

This study determined effective features in decision tree, neural network and regression models. Since the goal is to measure the effect of features in different models for each of the six categories and also to determine the most important feature for each section based on results of the three models used, the results of three algorithms are compared with each other.

These algorithms are selected because of the type of data, high accuracy and simplicity; moreover, they are the most practical algorithms, which have not been used simultaneously in similar studies, except for different variables and for obtaining results other than objectives of this study.

The decision tree CART produces a decision tree which tends to predict and classify future observations and can be used for continuous and discrete variables.

Neural networks are used for classification, clustering, prediction and pattern recognition; this method is preferred because of its high reception capability for noisy data and high accuracy in data mining.

Using the information of previous years and through supervisory methods, growth rate of electricity consumption is examined in the coming years. Then comparisons are made with the amount of power consumed in previous years. Eventually, the amount of electricity required will be achieved in the coming years. The type of classification in terms of the timing used in this study is long-term.

The following objectives are pursued in this study:

- Evaluation of data mining techniques and software required for this study
- Development of a suitable model for classifying subscribers based on power consumption
- Classification of consumption using the developed model
- Provision of a research field for modifying electricity consumption to minimize the problems associated with high power consumption

The collected data was preprocessed. Preprocessing output was considered as learning input. Note that learning input data was divided into two parts: training and test. In the learning phase, the models were applied on the test data; the results obtained by predicting test data were used to determine accuracy of the prediction.

Considering the attempt to prove validity of the suggested model, evaluation criteria were applied to determine validity of the algorithms and determine the best algorithm and the entire framework of the suggested model was repeated by changing the type of the target feature from continuous to discrete. This repetition tends to prove validity of the suggested model by comparing the results obtained by running the algorithms used for continuous and discrete values of the target feature.

2 MATERIALS AND METHODS

This section first describes the dataset used. Then, the software used with the method for developing the prediction model is explained.

2.1 Techniques and Algorithms Used for Subscriber Classification

There are various methods and algorithms for classifying subscribers, including the decision tree CART, neural network and regression, which are used in this study. Several software applications are currently provided by reputable software companies for data mining. Clementine software was used to implement data mining algorithms in this thesis.

2.2 Data Classification into Training and Testing Data

After implementing the data preprocessing steps, 75% of the data was used for training and 25% was used for testing. Data was randomly divided by the software used for data mining. In terms of the number of data for each set, the number of training data was considered to be greater than the test data. In this study, 75% of all data was considered for training and 25% was considered as testing data, which was done in the software by the partition node.

2.3 Executive Structure

In this section, the model and relevant analyses are developed based on available data using the Clementine software. According to the type of data, CART algorithm, neural network and regression were selected. Data mining steps and results of implementation are shown in Fig. 1.

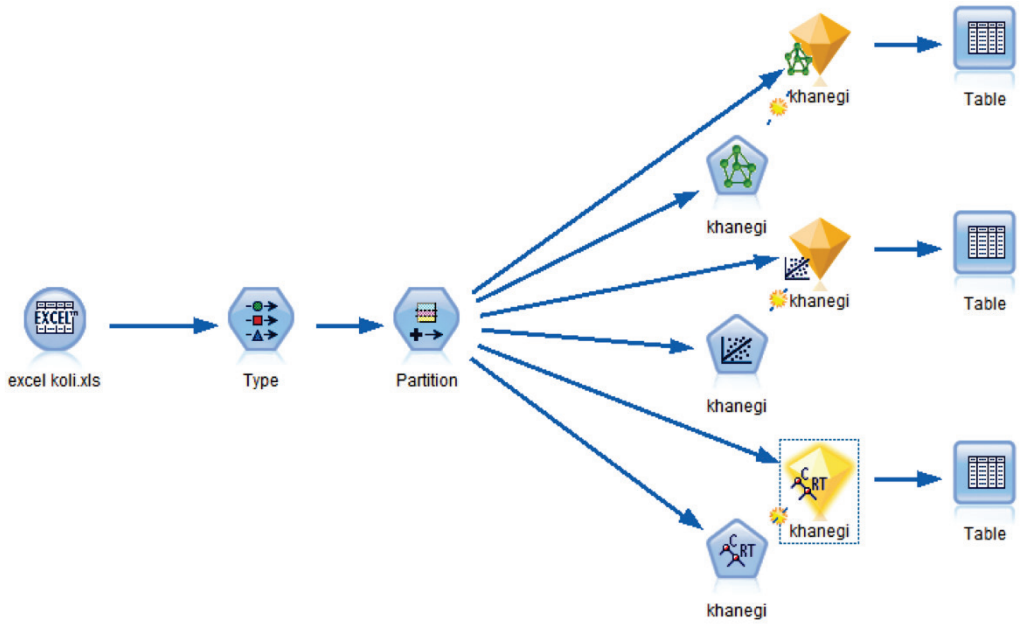

Figure 1 General instructions for algorithms used for various consumptions

Excel node: to insert the database into the software
Type node: to determine the type of attributes
Partition node: to determine the number of training and testing records
Table node: to display all database records
Table node: to allow the user to view the data table

3 IMPLEMENTATION OF ALGORITHMS

3.1 Implementation for Domestic Consumptions

3.1.1 Implementation by Decision Tree CART

One of the most commonly used methods for predicting is decision tree. The decision tree CART generates a decision tree which tends to predict and classify future observations. CART can be used for continuous and discrete variables. In the continuous run, the target variable is continuous. In the generated tree, all data is located in the root node at the highest point. Based on the variable which can provide the highest homogeneity for each branch, the root is branched. This continues until it reaches the data in each node which has the most possible homogeneity. Effective features of the CART algorithm are shown in Tab. 1 for domestic consumptions.

Table 1 Effectiveness of features of CART algorithm for domestic consumption

Effectiveness	
Population	0.92
Humidity	0.04
Temperature	0.04

3.1.2 Implementation by Neural Network

After normalizing, the data was divided into training and testing groups; using MLP option of the Method tab, all settings and network attributes were considered with the proper topology. In this case, lower number of hidden layers which can train the network faster and generalize better was selected. Tab. 2 shows result of the neural network topology and effectiveness of features for domestic consumptions.

Table 2 Effectiveness of features of neural network for domestic consumption

Effectiveness	
Population	0.79
Humidity	0.09
Temperature	0.12

3.1.3 Implementation by Regression

After normalizing the data, the data was divided into training and testing groups. Using Enter in the Method tab, *point* = 0.05 was considered as the best attribute of the regression method. Tab. 3 shows effective features of regression for domestic consumptions.

Table 3 Effectiveness of features of regression model for domestic consumption

Effectiveness	
Population	0.93
Humidity	0.05
Temperature	0.02

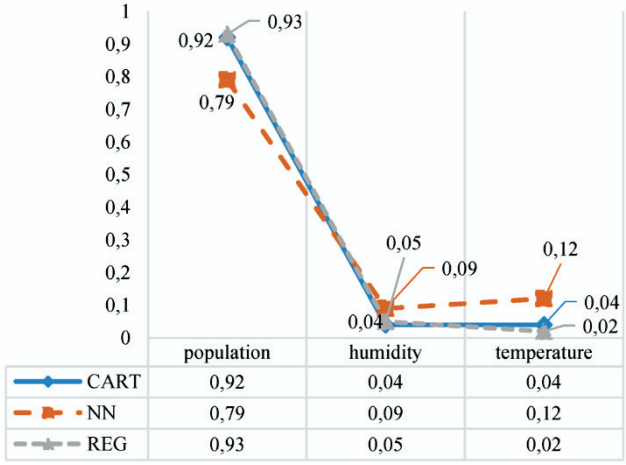

Figure 2 Results of test by regression model, neural network and decision tree CART for domestic consumptions

3.1.4 Results of Three Models for Domestic Consumptions

Given that the goal is to measure the effect of features in different models for each of the six classes (which is domestic consumption here) and to determine the most

important feature for each class based on the results of three models used, the results of three algorithms are compared; the results of this comparison are shown in Fig. 2. As results of three models show, the most important feature in domestic consumption is population.

3.2 Implementation for Agricultural Consumptions

3.2.1 Implementation by Decision Tree CART

Effective features of the CART algorithm are shown in Tab. 4 for agricultural consumptions.

Table 4 Effectiveness of features of CART algorithm for agricultural consumption

	Effectiveness
Population	0.87
Humidity	0.08
Temperature	0.05

3.2.2 Implementation by Neural Network

Tab. 5 presents effective features for agricultural consumptions.

Table 5 Effectiveness of features of neural network for agricultural consumption

	Effectiveness
Population	0.90
Humidity	0.07
Temperature	0.03

3.2.3 Implementation by Regression

Tab. 6 shows effective features of regression for agricultural consumptions.

Table 6 Effectiveness of features of regression model for agricultural consumption

	Effectiveness
Population	0.92
Humidity	0.04
Temperature	0.04

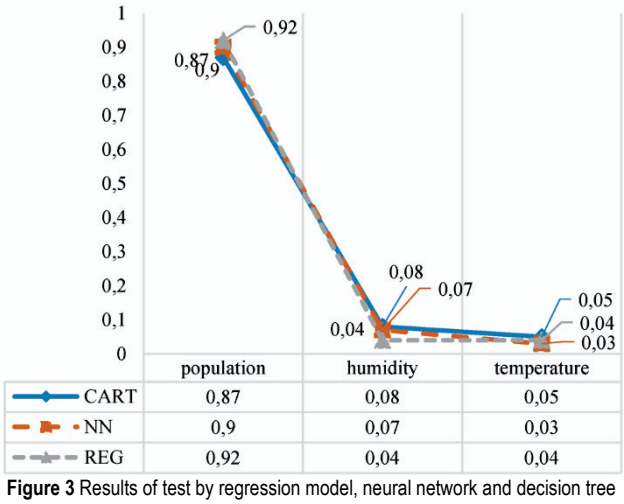

Figure 3 Results of test by regression model, neural network and decision tree CART for agricultural consumptions

3.2.4 Results of Three Models for Domestic Consumptions

Given that the goal is to measure the effect of features in different models for agricultural consumption and to determine the most important feature for each class based on the results of three models used, the results of three algorithms are compared; the results of this comparison are shown in Fig. 3. As results of three models show, the most important feature in agricultural consumption is population, followed by humidity and temperature, respectively.

3.3 Implementation for Industrial Consumptions

Effective features of decision tree, neural network and regression were determined for domestic consumption. The results of this comparison are shown in Fig. 4. As results of three models show, the most important feature in industrial consumption is population, followed by humidity and temperature, respectively.

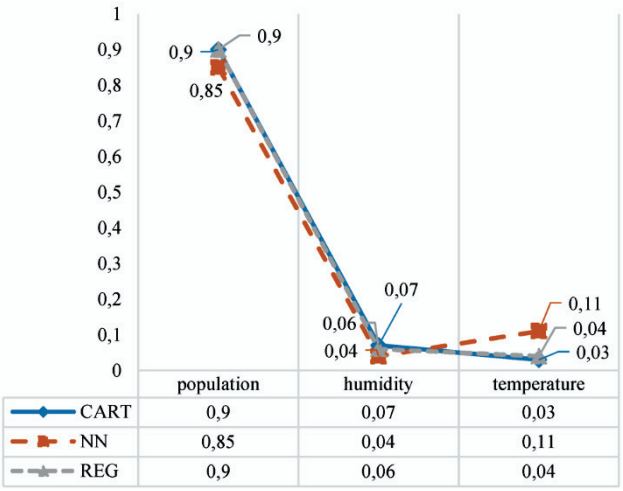

Figure 4 Results of test by regression model, neural network and decision tree CART for industrial consumptions

3.4 Implementation for Commercial Consumptions

Effective features of decision tree, neural network and regression were determined for commercial consumption. The results of this comparison are shown in Fig. 5. As results of three models show, the most important feature in commercial consumption is population, followed by humidity and temperature, respectively.

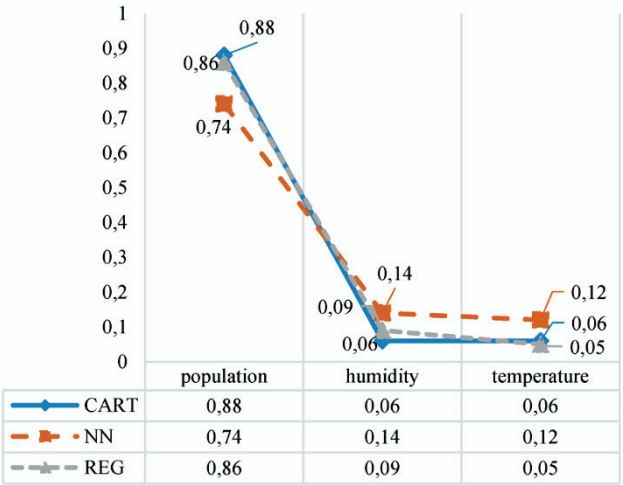

Figure 5 Results of test by regression model, neural network and decision tree CART for commercial consumptions

3.5 Implementation for Roads

Effective features of decision tree, neural network and regression were determined for roads. The results of this comparison are shown in Fig. 6. As results of three models show, the most important feature in roads is population, followed by humidity and temperature, respectively.

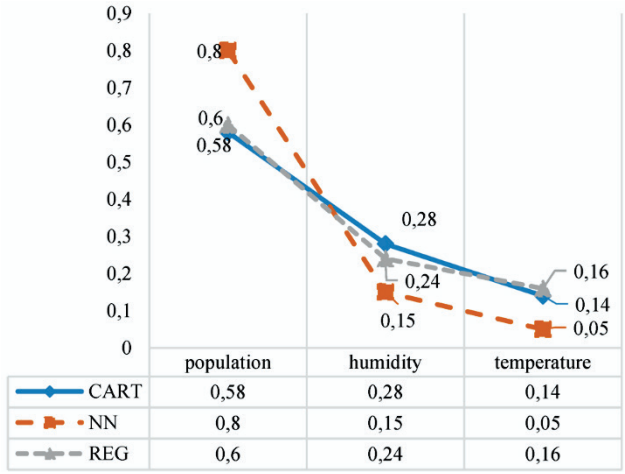

Figure 6 Results of test by regression model, neural network and decision tree CART for roads

3.6 Implementation for General Consumption

Effective features of decision tree, neural network and regression were determined for general consumption. The results of this comparison are shown in Fig. 7. As results of three models show, the most important feature in general consumption is population, followed by humidity and temperature, respectively.

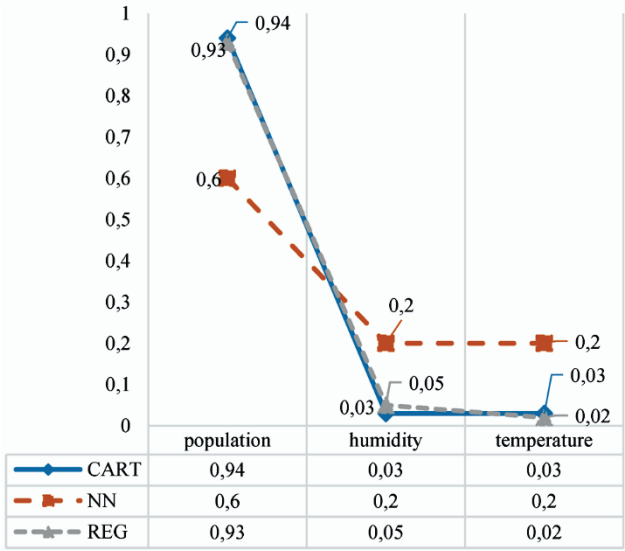

Figure 7 Results of test by regression model, neural network and decision tree CART for general consumption

4 DISCUSSION

According to data mining process, the data was collected, preprocessed and prepared as inputs of the learning models. The algorithms used were run on the dataset and the results indicate that population had the highest effect on the amount of power consumed in each of the six classes. This is consistent with electricity experts, because the amount of electricity consumed (not steadily) is likely to increase when the number of subscribers of each class increases. The second effective feature of power consumption in six classes is humidity, which in many classes has a relatively equivalent effect with the effect of temperature on power consumption. Effectiveness of population, humidity and temperature on six classes and the run algorithms are shown in Tab. 7.

Table 7 Effectiveness of features in regression model, neural network and decision tree CART for six classes of consumption

Class	Algorithm			Feature
	CART	NN	REG	
Domestic	0.92	0.79	0.93	Population
	0.04	0.09	0.05	Humidity
	0.04	0.12	0.02	Temperature
Agricultural	0.87	0.9	0.92	Population
	0.08	0.07	0.04	Humidity
	0.05	0.03	0.04	Temperature
Industrial	0.9	0.85	0.9	Population
	0.07	0.04	0.06	Humidity
	0.03	0.11	0.04	Temperature
Commercial	0.88	0.74	0.86	Population
	0.06	0.14	0.09	Humidity
	0.06	0.12	0.05	Temperature
Roads	0.58	0.8	0.6	Population
	0.28	0.15	0.24	Humidity
	0.14	0.05	0.16	Temperature
General	0.94	0.6	0.93	Population
	0.03	0.2	0.05	Humidity
	0.03	0.2	0.02	Temperature

In Tab. 7, maximum effect of various features for each of the algorithms is determined by different colors. In the decision tree CART, the highest effect on consumption is related to general consumption for population (0.94) and roads for humidity (0.28) and temperature (0.14). In the neural network, the highest effect on consumption is related to agricultural consumption for population (0.90) and general consumption for humidity (0.20) and temperature (0.20). In the regression model, the highest effect on consumption is related to general and domestic consumptions for population (0.93) and roads for humidity (0.24) and temperature (0.16).

5 CONCLUSION

Excessive electricity consumption is one of the major problems that all distribution companies are involved with. Some of them try to resolve this problem by measures such as advertising and buying new low power devices. This has highlighted the prediction of outcome of subscriber classification methods for both consumers and subscribers and managers and decision makers. For this reason, this study analyzed the data in this area to describe power consumption, which is the same classification of electricity subscribers. The results of this study will help managers in the field of power industry in their respective decisions and officials of the Ministry of Energy in Iran to pursue macroeconomic policies to increase electricity consumption. In the previous sections of this study, the framework was presented; this framework includes the steps to collect data on monthly electricity consumption (domestic, commercial, etc.); prepare, pre-process and normalize data; select important features in predicting

daily load; divide data into training and testing sets; run CART decision tree algorithms, neural network and regression; evaluate algorithms; and finally compare the accuracy of algorithms and choose the best algorithm. The work was based on the main framework and acceptable results were obtained.

After collecting the required data, data was pre-processed. Output of the preprocessing step was considered as input of the learning section. Note that the input data of learning model was divided into two parts: training and testing. In the learning phase of the algorithms, the developed models were run on the test data and the results of testing data prediction were used to determine accuracy of the prediction.

To improve performance of algorithms and future works, the following are suggested:

- Collect data extensively from several distribution companies with almost identical climates
- Provide a complete database for importing all features in the power distribution company
- Design a decision support software to help managers to predict the level of power consumption and classify subscribers
- Collect more influential fields which can affect the outcome of this thesis and give the list of these features to managers to receive information from power distribution companies
- Determine the sensitivity and specificity of each algorithm and compare it with each other
- Use combined algorithms: comparing the optimized algorithm using other algorithms (fuzzy logic, SVM, etc.)
- Use the suggested models to improve energy analysis in different sectors (water, gas, etc.) according to the specific energy feature.

6 REFERENCES

[1] Motameni, A., Jafari, E., & Mojarad, F. (n. d.) *Customer relation management*. s. l.: Bazargani Press.

[2] Taherpour-Kalantari, H. & Tayebi, T. A. (2010). The relationship between customer relation management and marketing performance. *Business Management Vision, 9*(1), 109-122.

[3] Hadizadeh-Moghadam, A., Raminmehr, H. & Haj-Moghani, R. (2010). Successful implementation model for customer relation management. s. l., s. n.

[4] Mirdehghan, S. & Saadatooo, F. (2014). Management and Optimization of Power Consumption Using Data Mining - A Case Study of Yazd Province. *National Conference on Energy Conservation in Science and Engineering*.

[5] Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. *Energy*, *34*(9), 1413-1421. https://doi.org/10.1016/j.energy.2009.06.034

[6] Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. *Energy*, *35*(2), 512-517. https://doi.org/10.1016/j.energy.2009.10.018

[7] Yan, Y. Y. (1998). Climate and residential electricity consumption in Hong Kong. *Energy*, *23*(1), 17-20. https://doi.org/10.1016/S0360-5442(97)00053-4

[8] Dent, I., Aickelin, U., & Rodden, T. (2011). Application of a clustering framework to UK domestic electricity data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2829232

[9] Dent, I., Aickelin, U., Rodden, T., & Craig, T. (2012). Finding the creatures of habit; Clustering households based on their flexibility in using electricity. https://doi.org/10.2139/ssrn.2828585

[10] Kavaklioglu, K., Ceylan, H., Ozturk, H. K., & Canyurt, O. E. (2009). Modeling and prediction of Turkey's electricity consumption using artificial neural networks. *Energy Conversion and Management*, *50*(11), 2719-2727. https://doi.org/10.1016/j.enconman.2009.06.016

[11] Li, D. C., Chang, C. J., Chen, C. C., & Chen, W. C. (2012). Forecasting short-term electricity consumption using the adaptive grey-based approach—An Asian case. *Omega*, *40*(6), 767-773. https://doi.org/10.1016/j.omega.2011.07.007

[12] Kaur, N. & Kaur, A. (2016). Predictive modelling approach to data mining for forecasting electricity consumption. In Cloud System and Big Data Engineering (Confluence), the 6th International Conference (pp. 331-336). IEEE. https://doi.org/10.1109/CONFLUENCE.2016.7508138

[13] Abolfazli, H., Asadzadeh, S. M., Nazari-Shirkouhi, S., Asadzadeh, S. M., & Rezaie, K. (2014). Forecasting rail transport petroleum consumption using an integrated model of autocorrelation functions-artificial neural network. *Acta Polytechnica Hungarica*, 11(2), 203-214. https://doi.org/10.12700/APH.11.02.2014.02.12

[14] Dalfard, V. M., Asli, M. N., Nazari-Shirkouhi, S., Sajadi, S. M., & Asadzadeh, S. M. (2013). Incorporating the effects of hike in energy prices into energy consumption forecasting: A fuzzy expert system. *Neural Computing and Applications, 23*(1), 153-169. https://doi.org/10.1007/s00521-012-1282-x

[15] Kuo, C. F. J., Lin, C. H., & Lee, M. H. (2018). Analyze the energy consumption characteristicsand affecting factors of Taiwan's convenience stores-using the big data mining approach. *Energy and Buildings*, 168, 120-136. https://doi.org/10.1016/j.enbuild.2018.03.021

[16] Mokhtari Sadaqati, M. R., Rezvani Khorashadizadeh, R., Ahmadi Darmian, E., & Akbari, M. (2011). Designing and Implementing an Intelligent Software System Based on Data Mining Techniques for Classifying Subscribers of South Khorasan Telecommunication Company. *Fifth Iran Data Mining Conference*, Amir Kabir University of Technology.

Contact information:

Elham PEYK
(Corresponding author)
Department of Computer, Faculties of Graduate Studies,
Rasht Branch, Islamic Azad University, Rasht, Iran
elham.peyk@hotmail.com

Asadollah SHAHBAHRAMI
Department of Computer Engineering, Faculty of Engineering,
P. O. Box: 3756-41635 University of Guilan, Rasht, Iran